How to involve structural modeling for cartographic object recognition tasks in high-resolution satellite images?

Guray Erus, Nicolas Lomenie

How to involve structural modeling for cartographic object recognition tasks in high-resolution satellite images?

Guray Erus and Nicolas Loménie

Department of Mathematics and Informatics, CRIP5 laboratory, SIP team, University Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France

Abstract

With the new generation of satellite systems, very high resolution satellite images will be available daily at a high delivery rate. The exploitation of such a huge amount of data will be made possible by the design of high performance analysis algorithms for decision making systems. In particular, the detection and recognition of complex man-made objects is a new challenge coming with this new level of resolution. In this study, we develop a system that recognizes such structured and compact objects like bridges or roundabouts. The original contribution of this work is the use of structural shape attributes in an appearance based statistical learning method framework leading to valuable recognition and false alarm rates. This hybrid structural/statistical approach aims to construct an intermediate step between the low-level image characteristics and high-level semantic concepts.

Key words: Object recognition, structural analysis, satellite images, CBIR

1 Introduction

The new satellite systems like PLEIADES or QuickBird will provide satellite images up to 70 cm per pixel spatial resolution with a high delivery rate. Thus, new challenges for automatic interpretation of these valuable images are coming up. Much work has been done so far about the segmentation/recognition of textured areas like urban ones at low resolution. The detection of transport

[1] Corresponding author.
Tel.:+1-334-787-75-11;
E-mail address: guray.erus@uphs.upenn.edu

networks like roads or rivers considered as lineic structure networks in images has also been intensely studied and efficient algorithms have been proposed so far like in (Tupin et al., 1998; Mayer et al. , 2006). However, few results have been presented for the recognition of complex, cartographic structures like bridges or roundabouts. These versatile objects do not lend themselves to classical statistical approaches for which a set of quantitative measures about the intensity distribution within a normalized window feeds a global statistical classifier used to discriminate between different objects. In the case of roundabouts for instance, the spatial configurations of object's parts are so versatile that a global statistical approach is not applicable. While the decomposition of the objects into parts is almost unavoidable, a purely structural part-based approach (Erus and Lomenie , 2005), is not appropriate to be applied on a real detection task, mainly due to the difficulty of segmenting the target objects to well-defined primitive shapes.

We propose in this study a hybrid approach based on learning the spatial configuration of structural primitives constituting an object, according to their statistical distribution in a labeled training set. Our method is evaluated on a cartographic object recognition task defined in the framework of a national program called Technovision that aims to assess the state of the art in object recognition (project ROBIN). For that academic-industrial joint study, The French Space Agency (CNES) prepared a database of satellite images containing cartographic objects. The recognition task consists of classifying images of size 100x100 pixels into one of the categories, like bridges, roundabouts, crossroads or isolated buildings.

In the object category recognition domain, the need for more structural analysis is progressively growing. In particular, low-level as well as high-level spatial relationships of object components are used as part of the object model definitions (Ferrari et al. , 2008). However, the handling of such spatial relationships is not obvious since symbolic/linguistic reasoning is more or less involved (Erus and Lomenie , 2005) and the concept of spatial ontology is not straightforwardly usable in current image processing lines (Hudelot , 2005). However preliminary examples of the use of spatial relations in a recognition task can be found for example in Colliot et al. (2006); Cao (2009).

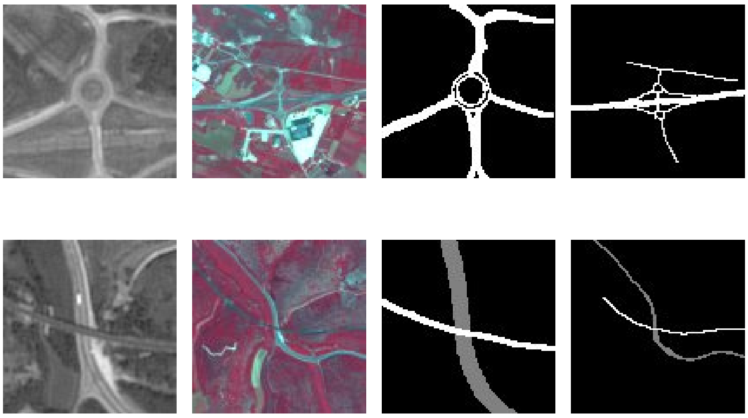
2 The image database

In the frame of the ORFEO program [2], CNES prepared a database of high resolution cartographic object images. These images are simulations of the

[2] The research program set up to prepare, accompany and promote the use and the exploitation of the images derived from future Pleiades satellites

PLEIADES acquisition system based on SPOT5 images. In fact, the actual PLEIADES images will be available around 2010. The images are 100 x 100 pixels windows within which the object of interest is approximatively centered. In each of the 10 defined categories, there are nearly 100 such sample images. For each object instance in each category, this database is made of (see Fig. 1) :

(1) the panchromatic image of the object;
(2) the multi-spectral image of the object;
(3) the manually segmented image as a mask drawn by an expert on both previous images.

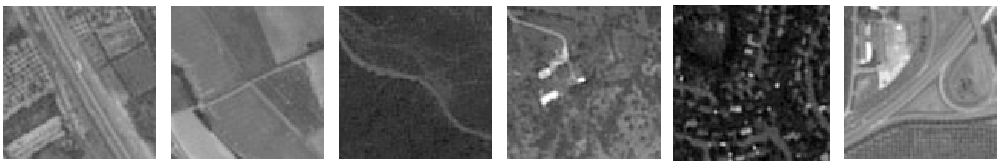


(a) (b) (c) (d)

Fig. 1. SPOT5 images of cartographic objects exhibiting structures, a roundabout and a bridge (a). The panchromatic images, (b). The multi-spectral images, (c)-(d). The corresponding manually segmented images

The panchromatic images are acquired at the resolution of 2.5 m. per pixel and the multi-spectral ones at 10 m. per pixel. The challenge of that study however is to restrict oneself to the panchromatic images to handle the case for which only this channel is available due to transmission rate issues for instance, but also to be as generic as possible. Obviously, the use of such complementary information, if available, should improve the performance of the operational recognition system dramatically.

In this study, we focus our attention to the classification of two cartographic object categories, *"Roundabouts (RA)"* and *"Bridges (BR)"*, which have a highly variable and compact structure. The long-term objective is the automatic detection of cartographic objects on very large satellite images (e.g. 24000 x 24000 pixels). However, we restrict ourself to the classification task proposed by the CNES, where the goal is to determine the categories of objects on manually extracted and labeled image patches. The basic premise is that a pre-processing module will focus the attention on the regions that may contain the target objects, and extract candidate regions. This initial coarse detection module can be constructed by edge density analysis or using one of the recent interest point detection algorithms for instance. The global recognition task (detection and classification) requires a sophisticated strategy and is still an active research issue. Figure 2 presents sample panchromatic images belonging to different object categories.



(a) (b) (c) (d) (e) (f)

Fig. 2. Panchromatic SPOT5 images of cartographic objects. (a). Highway, (b). Secondary road, (c). Path, (d). Isolated building (e). Housing estate, (f). Crossroad.

3 State of the art

Few works deal with the recognition task of such specific, highly structured objects in satellite or even in aerial images (Trias-Sanz and Loménie , 2003). Iqbal and Aggarwal (2002) proposed to use the perceptual grouping of edge features for classification. The main idea here is to count evidences of the target class in the image. The evidences are detected starting with the line segments and applying grouping rules to obtain, hierarchically, co-terminations, L and U junctions, parallel lines and groups, and closed polygons. From these, three scalar features are calculated and the images are labeled as *Structure*, *Non-structure* or *Intermediate* using a nearest neighbor classifier. The important limitation of this method is that the low dimension of the final feature vector is only adequate for a very coarse classification.

Inglada (2007) developed a method to classify cartographic objects in satellite images. Low level and high level geometric descriptors are calculated from the image for learning an SVM classifier. Our method follows a similar approach in that our feature vector is obtained from a set of structural primitives and a supervised learning approach is used for the classification.

In recent years, *appearance based* approaches have been widely and successfully used in generic object detection problems (Agarwal and Roth , 2002; Csurka et al. , 2004; Dorkó and Schmid , 2003; Fei-Fei and Pietro , 2005; Felzenszwalb and Huttenlocher , 2005; Fergus et al. , 2003; Heisele et al., 2001; Leibe et al. , 2004; Opelt et al.; Shotton et al. , 2007). In this approach, the object is considered to be represented by its appearances from different viewpoints, and this representation is learned from a set of object images. As the global appearance of an object is highly variable, the recent methods use in general local descriptors which remain stable on different instances of the object. The object consists of a collection of its parts, and the objective of the learning problem is to learn the characteristic parts of an object and their spatial organization. This approach is also conform to the neuro-psychophysical hypothesis of Biederman (1987) (and more recently Biederman (2007)), according to which an object is composed of a small set of elementary geometric shapes called *geons* and the first stage of human object recognition is the recognition of *geons*. The definition of the basic components and the modeling of their spatial arrangement is performed using various methods, giving birth to several detection algorithms.

A very common approach in extracting object parts is to detect local interest points on the image using detectors like Harris, maximally stable regions, difference of Gaussians, or the popular SIFT method. In Viola and Jones (2004) a very large number of simple Haar-like rectangular regions are extracted as features. Shotton et al. (2007) proposed to use contour fragments from the outer boundary of the objects. In Ferrari et al. (2008) contour segments are used as descriptors.

Spatial relations between parts can be modeled in a large spectrum of approaches, from appearance-only models that ignore them completely as in the *bag of words* approach (Csurka et al. , 2004; Fei-Fei and Pietro , 2005), to structural, graph-based approaches as *pictorial structures model*, or *constellation model*. The *Pictorial Structures* proposed initially by Fischler and Elschlager (1973) for face recognition has been recently used for the moving person recognition task in Felzenszwalb and Huttenlocher (2005). The idea is to represent the configuration of the components by elastic connections. The recognition proceeds by the minimization of a cost function assessing the needed deformation of the elastic structural model to the unknown configuration. By an expectation-minimization approach, Weber et al. (2000) learn a *constellation model* which is a star-shaped graph representing the components and their position around a fixed point, often chosen as the center of the object.

Obviously, these approaches rely dangerously on the quality of the component segmentation process. We can notice here the important ambiguous role played by the spatial reference. It is often the cornerstone for the structural modeling. In medical images, the anatomical substructure provides a straightforward spatial reference (left and right lung for instance). Contrarily, in the case of aerial images, one loses any classical spatial reference like above, below, left or right. A central point is often the easiest choice to be made. But, according to us, the choice of a spatial reference should deserve much more investigation.

In Leibe et al. (2004) a codebook of local appearances is obtained by clustering rectangular image patches according to their normalized greyscale correlation measure. An implicite shape model, in the form of the spatial probability distribution of each cluster relative to the object center, is constructed using a probabilistic voting procedure, by matching patches extracted from the training set to the clusters. A similar approach in the construction of the codebook is used in Agarwal and Roth (2002), but binary spatial relations are represented by a sparse matrix obtained from the histogram of distances and angles all pairs of patches. We can finally cite the work of Fergus et al. (2003) where the components and the structure are learned jointly using EM algorithm.

The whole set of these methods that attempt to learn an object category from the appearance of local components and their spatial organization, offers a methodological framework well adapted to cartographic objects. However they have been used so far for the recognition of object categories exhibiting little structural variation like bicycles, cars or faces. Furthermore, the target objects have generally a constant viewpoint and orientation, according to which the spatial relations are defined. By nature, the configuration of a roundabout or a bridge is much more versatile. We believe that, in the case of cartographic objects, the parts of objects can be naturally associated to geometric primitives with a still higher level of representation. At the new resolution of satellite images, that representation is totally justified by efficient low level image processing methods to extract such features.

In satellite imagery, most of the accomplished studies focus on the extraction of extended surfaces like road networks or urban areas (Lorette et al. , 2000; Tupin et al., 1998). This specific topic of interest was related with the low level of resolution at which the objects of interest were much more handled as textured areas to be segmented in the image. Now, the resolution makes it possible to focus on more localized and compact objects like a roundabout or a bridge.

To end up with the structural pattern recognition community, a few works cope with the purely structural modeling of complex objects in order to address the difficult issue of the semantic gap between the low-level descriptors and the high-level concepts. The Attributed Relational Graphs (ARG) are often used as a high-level model of representation for the arrangement of regions (Petrakis and Faloutsos , 1995; Shao and Kittler , 1999) or the modeling of skeletons (Bardinet et al. , 2000; Di Ruberto , 2004). The construction or learning of such structural models is addressed in Hong and Huang (2004); Sangineto (2003); Cordella et al. (2002). Even though these purely structural approaches are very interesting, they are not efficient as such for recognition/detection purposes in cluttered environments where the object of interest and the background are not easily separable.

4 The cartographic codebook modeling

The component-based approaches are well adapted for the modeling of cartographic objects with two main reasons:

- the articulated nature of the objects: for example, a roundabout is composed of a central circle and roads articulated around the circle;
- the intra-class variation is very high and this variation is often due to affine transformations of the object components.

Cartographic objects have well-defined, mostly geometric structures. For this reason, we propose to exploit the geometric nature of cartographic objects in a component-based learning framework. The originality of this approach lies in:

- the intensive use of geometric primitives to build the appearance codebook;
- the joint learning of this cartographic codebook and of the structure.

In our work, the object components correspond to geometric primitives extracted from the images. However, due to their genericity, their informative content is weak and the matching between two such primitives in two images is not a definitive clue about the similarity between these two images. This is the reason why the joint modeling of components and structure is required. We propose to extract all eventual geometric primitives that may belong to the object, and learn jointly the components and the structure by defining the components not only by their geometric properties but by their spatial properties as well. Then, we select the most significant spatio-geometric primitives to represent the object. A simplified structural representation of this object may then be displayed for the naive user of the final interface.

Let $\mathbf{P} = \{p^1 \dots p^m\}$ where $\{p^i \in \Re^{d_i}\}_{i=1}^m$ be the set of primitive types, each represented by a corresponding feature vector denoting attributes such as position, orientation and geometric properties. Let $\mathbf{I_{Train}}$ be the training set consisting of images containing the target object with label l. Primitives are extracted from each image in $\mathbf{I_{Train}}$ and combined in $\mathbf{p_{Train}} = \{p_1, \dots, p_n\}$ where $\{p_i \in \mathbf{P}\}_{i=1}^n$. A codebook dictionary is built up by grouping the primitives in $\mathbf{p_{Train}}$ relatively to their types $i \in \{1, \dots, m\}$, and clustering them in the corresponding feature space $\Re^{d_i}$. Then the best clusters, which correspond to words in our structural codebook, are automatically selected. A codebook is built up for each category of object. This codebook, which we call as the Structural Model Codebook (*SMC*), works as a structural model carrying both geometric component and structure information. Note that the learning step does not need negative examples and is thus apparented to the original category of one-class classifiers (D. M. J. Tax , 2001; Wang, Q. et al. , 2004).

In a probabilistic framework, the *SMC* can be considered as a model that is used for estimating $p(p^i|l), i = \{1, \ldots, m\}$, the multivariate probability density function (**pdf**) of each primitive type i defined on $\Re^{d_i}$, given the object category l.

Given a test image I, from which a set of primitives $\mathbf{p_I} = \{p_1, \ldots, p_q\}$ is obtained, the classification into category l is essentially performed by calculating a class membership score obtained by accumulating the evidences of observing a component of the object category with label l in $\mathbf{p_I}$. The likelihood of a primitive to belong to category l is estimated from the **pdf**s learned by the *SMC*.

Algorithm 1 briefly presents an overview of the algorithmic procedure. A detailed formulation of each module is given in the following subsections.

Algorithm 1 Global Algorithm of the components-and-structure method

Require:
  $\mathbf{I_{Train}}$: Learning set of positive images
  $\mathbf{I_{Test}}$: Test set of images
Ensure:
  $\mathbf{s}$: Membership scores of test images

  LEARNING
  for all $I_i \in \mathbf{I_{Train}}$ **do**
    Extract $\mathbf{p_i}$: the geometric primitives extracted from I_i
    Compute $\mathbf{a_i}$: the attribute vectors calculated from $\mathbf{p_i}$
  end for
  Construct $\mathbf{C_{object}}$: *clusters* obtained by clustering $\mathbf{a} = \bigcup \mathbf{a_i}$ by mean-shift clustering
  Construct $\mathbf{SMC_{object}}$, the structural model codebook of the object: *clusters* selected from $\mathbf{C_{object}}$

  CLASSIFICATION
  for all $I_i \in \mathbf{I_{Test}}$ **do**
    Extract $\mathbf{p_i}$: the geometric primitives extracted from I_i
    Compute $\mathbf{a_i}$: the attribute vectors calculated from $\mathbf{p_i}$
    for all $a_{ij} \in \mathbf{a_i}$ **do**
      $s_{a_{ij}}$: the likelihood of the attribute calculated from $\mathbf{SMC_{object}}$
    end for
    s_i : *normalized* $\sum s_{a_{ij}}$
  end for
  Return $\mathbf{s} = \bigcup \mathbf{s_i}$

4.1 The primitive extraction and representation

The segmentation of aerial images is not straightforward due to:

(1) the non-homogeneity of the radiometry of objects;
(2) the non-separability of the object from the background even for the human eye.

The manually segmented images of Figure 1.c-d illustrate the important amount of contextual information involved for the human expert to produce the segmentation ground truth. For this reason, we propose to extract geometric primitives that represent well the structure of man-made cartographic objects. The extraction is performed on the whole image, and eventual false positives are eliminated by the feature selection mechanisms at subsequent steps. Two types of primitives, *the straight lines*, and *the circle arcs* are extracted using both an edge-based and a region-based method. The two sets of primitives are used in conjunction in order to rely on a larger feature vector. The point is always to find a balance between the degree of expressiveness of a primitive and the difficulty to extract it. The more expressive the primitive is, the smaller the feature vector is, but more difficult the extraction becomes. To compare with Iqbal and Aggarwal (2002), circular primitives are also extracted, and a larger feature vector, taking into consideration primitives' geometric, relational and spatial attributes, is constructed.

The edge-based extraction is done on a Gaussian image pyramid with 4 levels, in order to detect the primitives on different scales. Given the original image I_0, the multiscale representation is obtained by

$$I_{l+1} = g * I_l, \;\; l \in \{0, 1, 2\}$$

where g is a Gaussian kernel and $*$ is the convolution operator. For each $\{I_i\}_{i=0}^{3}$ the following steps for the primitive extraction are applied and the resultant primitives are grouped together:

- Sub-pixel edge detection (Devernay , 1995) using a modified Canny (Canny , 1986) edge detector. Sub-pixel precision is necessary in order to guarantee a robust approximation by line segments or arcs.
- Grouping of edge points E_i into edge-chains $C_i = \{C_i^1, \ldots, C_i^p\}$, such that C_i^j is a list of connected edge pixels. A new chain is started in each junction.
- Polygonal approximation on each edge chain C_i^j using Douglas-Peucker algorithm (Douglas and Peucker , 1973), to obtain S_i^j, a set of adjacent line segments that fit C_i^j with the smallest error.
- Detection of arcs in each S_i^j. As a result of the polygonal approximation, the circular lines are also approximated by line segments. In order to de-

tect them, we propose a recursive fusion algorithm. For each adjacent line segment pairs $s_k, s_{k+1} \in S_i^j$, the circle c_k that best fits the corresponding edge points in a least squares sense is calculated, together with the approximation error $e(c_k, s_k, s_{k+1})$. If $e(c_k, s_k, s_{k+1}) < t$, a predefined circularity threshold, s_k and s_{k+1} are replaced by c_k. This procedure is repeated until $e(c_k, s_k, s_{k+1}) > t, \forall k$.

In parallel to edge-based extraction, region-based primitives are extracted by *mean-shift segmentation* (Comaniciu and Meer , 2002), an efficient nonparametric segmentation method. It's a generalization of the mean-shift clustering algorithm for segmentation, by mapping image pixels to a joint spatial-range domain, where for grey-level images the range corresponds to pixel intensity. Each pixel is associated with a significant mode of the joint domain density located in a predefined spatial and intensity neighborhood. We used a search window of size 6 in the intensity domain and of size 15 in the spatial domain.

This procedure generally results in an over-segmentation of the image that does not allow the detection of object components in different sizes. To overcome this problem, we applied an iterative fusion algorithm similar to the one applied for the line segments.

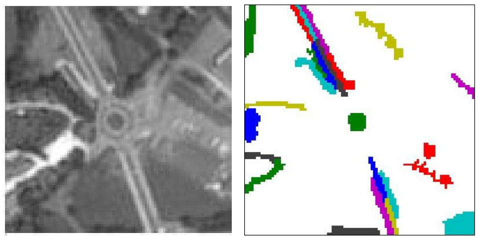
Let $\mathbf{R^0} = \{r_1, \ldots, r_n\}$ be the set of regions obtained applying the *mean-shift segmentation* on image I, $\mathbf{V^0} = \{v_1, \ldots, v_n\}$ the areas of these regions, $\mathbf{G^0} = \{g_1, \ldots, g_n\}$ their average intensities, and $\mathbf{N} \in \{0,1\}^{n \times n}$ the adjacency matrix of $\mathbf{R^0}$. A similarity matrix $\mathbf{S}$ is calculated as

$$S_{i,j} = \begin{cases} |g_i - g_j| & \text{if } N_{i,j} = 1; \\ +\infty & \text{if } N_{i,j} = 0. \end{cases}$$

A more sophisticated metric, that also considers region regularity for instance may also be used, but not preferred for efficiency reasons. At each iteration t the most similar two regions r_k and r_l are replaced by a new region r_{n+t} obtained by merging r_k and r_l. The intensity of the new region is calculated as

$$g_{n+t} = \frac{g_k \times v_k + g_l \times v_l}{v_k + v_l}$$

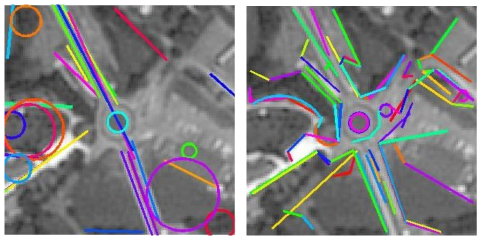
This procedure is repeated t^* times until the total number of regions is lower than a fixed threshold. For each region $r_i, i = \{1, \ldots, n + t^*\}$, a circularity score s_c^i and an eccentricity score s_e^i are calculated. s_e^i is defined as the scalar that specifies the eccentricity of the ellipse that has the same second-moments as r_i. Regions with high s_c are selected as circular primitives and those with high s_e are selected as line segment primitives. Figure 3 shows the regions selected before the fusion step.



(a) (b)

Fig. 3. The selected regions after the initial segmentation by Mean-Shift. (a). Original Image. (b) The linear and circular regions before the fusion step.

Figure 4 shows the primitives extracted from a roundabout image using the edge-based method and the region-based method.



(a) (b)

Fig. 4. The extracted primitives on a roundabout image (a) using the edge-based method, (b) using the region-based method

A set of geometric and spatial characteristics is associated to each extracted primitive. Representing spatial relationships by a set of numerical features is a challenging problem. We should note that the definition of the spatial reference is an important step for any attempt to model spatial relationships (Colliot et al. , 2006; Cao , 2009). In our case, due to the nature of the viewpoint, and because the objects of interest are centered in the image, the spatial reference is naturally selected as the central point in the image. The distance and the angle with respect to this reference point are used to represent the spatial properties of object components.

By the way, this representation around the center of the image enforces the invariance to rotation. The primitives are represented by two feature vectors f_{Circle} and $f_{Segment}$ (see Figure 5) :

- $f_{Circle} = \{d_C, r_C\}$, d_C = the distance between the center of the circle and the center of the image, r_C = the radius of the circle;

- $f_{Segment} = \{d_S, \theta_S, l_S\}$, $d_S =$ the distance between the center of the segment and the center of the image, $\theta_S =$ the angle between the segment and the line joining the center of the image to the farthest segment extremity from this center, $l_S =$ the length of the segment.

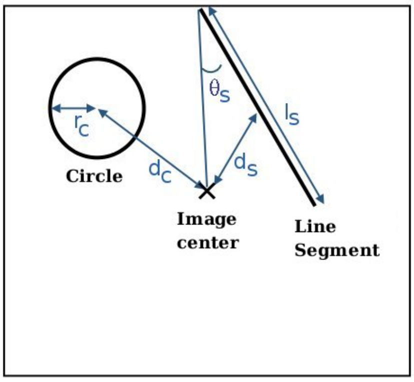


Fig. 5. Construction of the feature vector from the primitives.

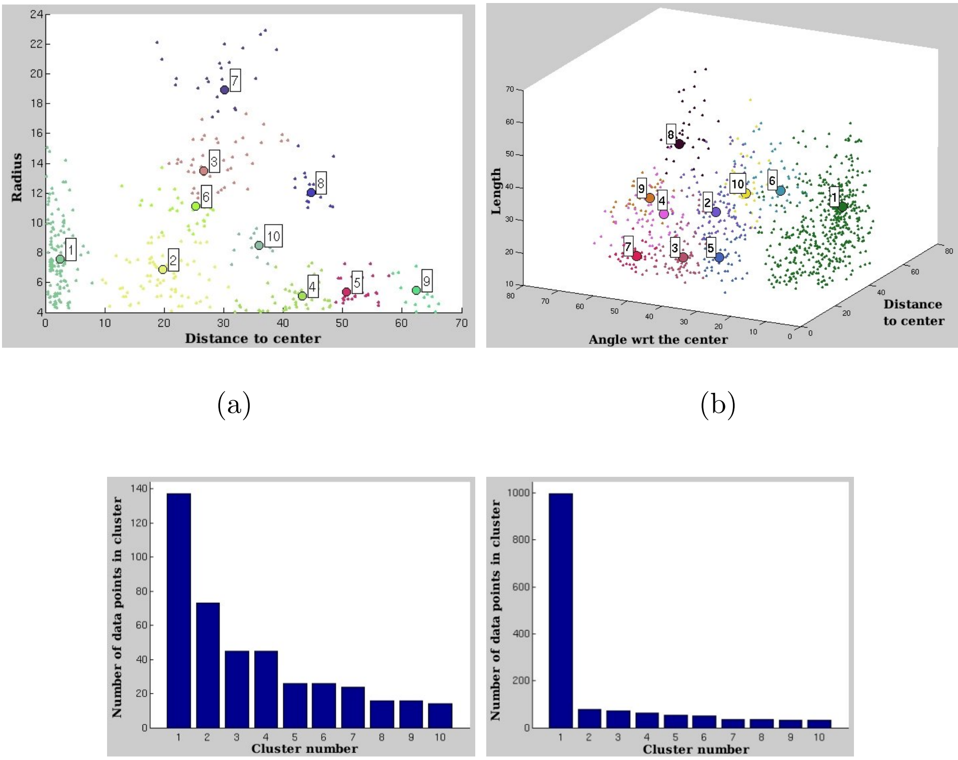
Notice that the parameter θ_S measures a certain radiality property of the segment primitive with respect to the center. We believe that a more formal analysis and definition of such spatial characteristics should be worth studying in the future for pattern recognition community. Anyway, in our study, this measure worked quite well to represent the spatio-structural configuration of interest.

4.2 Clustering

In Erus and Lomenie (2007) a bag of words approach is used, where a feature vector is constructed by accumulating the evidences of geometric, relational and spatial properties of the primitives. To do that, the attribute values of primitives are divided in bins, and the number of primitives in each bin is counted. The feature vector consisted of the concatenation of the values in each bin. An *Adaboost classifier* is used to select the most relevant features to describe the object. One weakness of this approach is the use of discrete intervals whose number and size is fixed empirically.

In this study, we propose to cluster primitives in the space defined by the values of their attributes using the *Mean-Shift clustering* algorithm (Cheng , 1995). Being a non-parametric clustering method that does not require to set a prior number of clusters and a prior shape for clusters, it is well adapted to our problem.

The *Mean-Shift clustering* algorithm proceeds by a gradient ascent procedure on the estimation of the local density in a fixed size window around each data point, until convergence. The stationary points obtained by this procedure



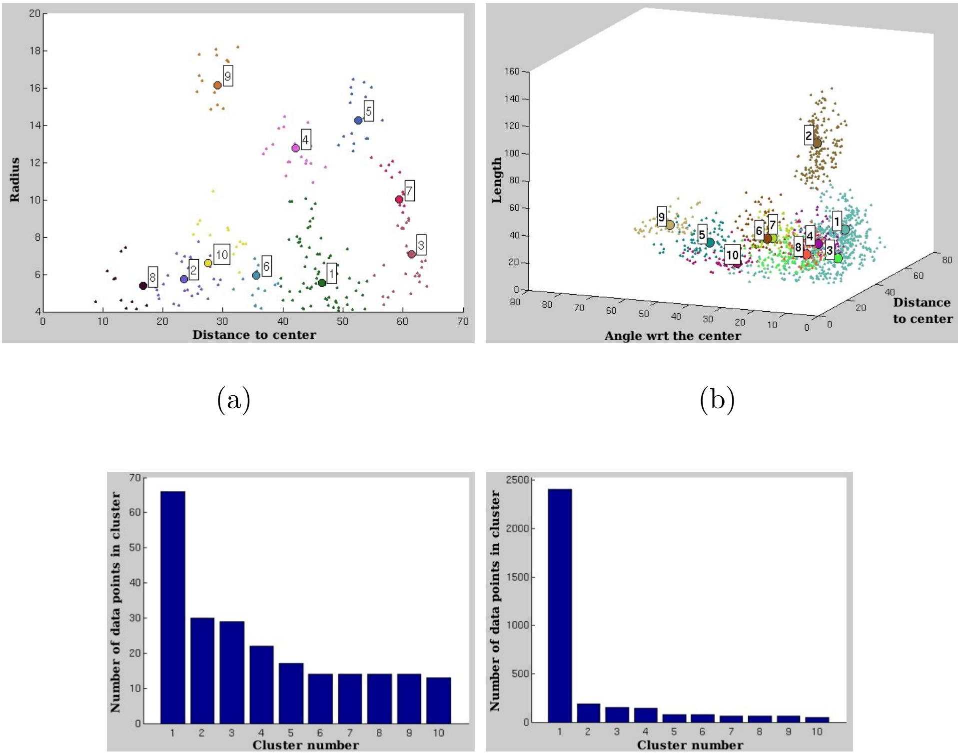
(a) (b)

(c) (d)

Fig. 6. Clustering of the primitives for the roundabouts. The data distribution and the first 10 largest clusters (a) for the circles, (b). for the line segments (the number of data points is reduced to obtain a better visibility). (c). The sorted number of points in the clusters for the circles, (d). for the line segments.

represent the modes of the distribution and the clusters are constructed by assigning each data point to a mode. The only parameter of the method is the normalized size $t \in [0, 1]$ of the window used to estimate the local density.

The mean-shift clustering algorithm is applied independently on line segment and circular primitives. The formers are clustered in a tri-dimensional space and the latters are clustered in a bi-dimensional space. The line segment primitives extracted from all training images belonging to the target object category are grouped together. Each feature vector $f_{Segment}$ (calculated from each primitive) is considered as a data point, and the clustering algorithm is applied on the set of all data points to detect the modes of the distribution. The same procedure is also applied for the circular primitives. The window size is set empirically to $t = 0.1$. Figure 6 and 7 illustrate the clusters obtained for line segments and circles, and the sorted number of data points in each cluster for both object categories.



(a) (b)

(c) (d)

Fig. 7. Clustering of the primitives for the bridges. The data distribution and the first 10 largest clusters (a) for the circles, (b). for the line segments (the number of data points is reduced to obtain a better visibility). (c). The sorted number of points in the clusters for the circles, (d). for the line segments.

4.3 Cluster selection

The selection of the significant clusters is based on a heuristic rule depending on the number p_i of points in a cluster i normalized by the number of images in the training set n. Let d be the density threshold, the cluster i is said to be significant if the average value $\bar{n}_i = p_i/n$ of the number of points per image in the cluster i is higher than d. We set $d = 2$, that corresponds to observe a primitive in a selected cluster 1 times in every image in average (the value of d is doubled considering that two sets of primitives are used together).

Table 1 indicates the selected clusters for both object categories together with the statistics of the features within each cluster.

We would like to underline the correspondence between the learned clusters and the description of the target object category at a symbolic level: Assuming that the clusters have a Gaussian distribution, and setting the membership threshold as 1 σ for each attribute, a roundabout, for example, may be de-

Class	Primitif	Cluster	Mean			Standard Dev.		
			μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
RA	Circle	1	2.36	7.80		1.50	2.34	
	Segment	1	33.72	8.65	30.48	8.37	6.87	10.58
BR	Segment	1	29.78	7.45	39.06	11.38	6.15	17.55
	Segment	2	8.24	3.27	119.59	5.12	2.60	16.60
	Segment	3	25.64	20.07	26.03	3.40	5.13	8.13
	Segment	4	46.38	21.01	24.12	6.14	4.15	6.96

Table 1
Statistics of the primitives in selected clusters. μ_i and σ_i values correspond respectively to d_C and r_C attributes for the roundabouts, and d_S, θ_S and l_S attributes for the bridges

scribed as follows:

- circles with a distance to the center between 0.86 and 3.86 pixels (that is between 2.15 and 9.66 meters) and a radius between 5.46 and 10.14 pixels (between 13.65 and 25.34 meters) and
- segments with a distance to the center between 25.35 and 42.09 pixels (between 63.37 and 105.23 meters), a relative angle to the center between 1.77 and 15.52 degrees and a length between 19.90 and 41.06 pixels (between 49.74 and 102.65 meters).

4.4 Classification

For each cluster k, the Mahalanobis distance of a primitive p to the cluster k is given by:

$$d_M^k(\mathbf{p}) = \sqrt{(\mathbf{p} - \mu)^T \ \mathbf{\Sigma}^{-1} \ (\mathbf{p} - \mu)}$$

where Σ is the covariance matrix of the cluster. The distance $d_M^k(\mathbf{p})$ is used as a probability measure of membership to the cluster k, to assign a primitive p to one of the clusters.

In that way, for each primitive p_i extracted from a test image I, a membership score s_i is calculated as:

$$s_i = min_{k \in SMC_{object}} d_M^k(p_i)$$

To calculate the membership score of an image to a specific object category, the scores of extracted primitives are added using a sigmoid-like activation function, similar to the activation function used in neural networks to calculate the output value of a neuron:

$$f_\lambda(p) = 2 - \frac{2}{1 + e^{-\lambda p}}$$

The activation function performs as a smooth threshold, such that the accumulation of the scores of primitives having a high distance to the closest cluster center would be negligible. The value of the parameter λ is set adaptively for each cluster according to the number of data points in that cluster.

The final membership score of an image to a specific codebook SMC_{object} modeling an object category is calculated as:

$$Score_{SMC_{object}}(I) = \sum_{p_i \in I} f_\lambda(p_i)$$

5 Results

We tested our method on the cartographic objects base of CNES consisting of 962 images belonging to 10 categories. From these images, 72 are roundabouts and 99 are bridges. The image set is divided into a training set and a test set, each containing approximatively half of the objects in each category. A codebook for the roundabouts ($\mathcal{SMC}_{\mathcal{RA}}$) and a codebook for bridges ($\mathcal{SMC}_{\mathcal{BR}}$) are constructed independently using the training images, considering the images belonging to the target category as positive, and all other images as negative. For each image in the testing set, membership scores to both classifiers are calculated and they are classified accordingly.

Figure 8 displays the ROC curves representing the classification results using $\mathcal{SMC}_{\mathcal{RA}}$ and $\mathcal{SMC}_{\mathcal{BR}}$. The area under the ROC curve (AUC) is **0.9699** and **0.8432** respectively.

In Table 2 the classification results for the optimal decision threshold (corresponding to the highest *f-measure* value) are given.

We obtained an *f-measure* similar to that obtained using the *Adaboost classifier* (Erus and Lomenie , 2007) for the classification of roundabouts. However the *f-measure* of the classifier for bridges is 25% higher. Among the 49 images having the best scores in $\mathcal{SMC}_{\mathcal{BR}}$ we get 31 bridges instead of 24 bridges in our previous study.

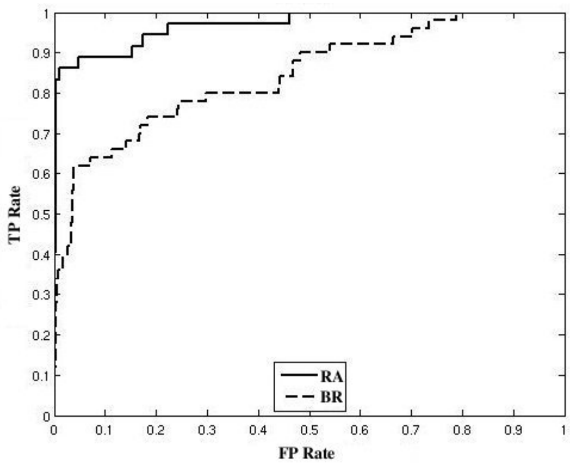


Fig. 8. The ROC curves associated to the classification by the $\mathcal{SMC}_{\mathcal{RA}}$ and the $\mathcal{SMC}_{\mathcal{BR}}$ classifiers.

Class	TP	FP	FN	TN	P	R	*f-measure*
Roundabout	30	3	6	442	0.909	0.833	0.870
Bridge	31	18	18	414	0.633	0.633	0.633

Table 2
The classification results obtained using an optimal threshold (TP=true positive, FP=false positive, FN=false negative, TN=true negative, P=precision, R=recall).

We applied random sub-sampling as a cross-validation strategy in order to test the robustness of the method. The data is split into two sets containing equal number of randomly selected positive and negative samples in all categories. The method is applied by using these two sets as training and testing sets. We repeated the random sub-sampling 100 times and calculated the mean and standard deviation (*std*) of the AUC for both object categories. We obtained a mean AUC of **0.9466** with *std* **0.0342** for the roundabouts, and a mean AUC of **0.838** with *std* **0.0251** for the bridges.

A qualitative analysis of the results is also performed by examining the categories that obtained high scores for each classifier, and observing the selected primitives for target objects with highest and lowest scores. As shown in Table 3, the categories *isolated buildings* and *crossroads* obtained high scores in $\mathcal{SMC}_{\mathcal{RA}}$. These are the categories which are the most similar to roundabouts structurally, and this result is mainly due to the small size of some of the roundabouts in the image base that does not allow to extract properly the geometric primitives. With higher resolution images we expect that the number of misclassified images would reduce significantly.

In $\mathcal{SMC}_{\mathcal{BR}}$ the non-bridge objects that obtained highest scores all belong to

Object category	Roundabout	Isolated building	Crossroad
Number of objects	30	4	2

Table 3
Number of objects that obtained the highest 36 scores in $\mathcal{SMC}_{\mathcal{RA}}$ grouped by their categories.

linear object categories (Table 4). These objects have all similar parts with bridges in different spatial configurations. Our model represents the spatial configuration of parts of objects implicitly using statistics of their geometric and spatial characteristics. The classification errors indicate the limits of this implicit modeling approach.

Object category	Bridge	Highway	Crossroad	National road	Railway
Number of objects	31	9	5	3	1

Table 4
Number of objects that obtained the highest 36 scores in $\mathcal{SMC}_{\mathcal{BR}}$ grouped by their categories.

Besides classification, the model allows us to visualize the structures of objects in target classes, by projecting the selected primitives on images. This visualization may be used by an interactive system, or as input to a subsequent processing system.

Figures 9 and 10 show the selected primitives for the roundabouts/bridges (using a gray level corresponding to their normalized score) that obtained the highest and lowest total scores. We observe particularly that the 9 bridges that obtained the lowest scores are all bridges on rivers. The rivers have in general a very irregular geometry that makes the extracted primitives very insufficient to represent the object.

6 Conclusion

In this study, we explored the use of high-level structural and spatial characteristics of man-made objects on satellite images in a statistical classification framework. Our method is based on recent appearance-based approaches, but is differentiated from other methods by the integration of structural primitives and their geometric and spatial attributes. We believe that the continuous technological developments in satellite imagery that makes available higher resolution images justify our quest for a structural approach, which is also applicable on practical real problems.

We obtained promising results that might be improved by the addition of intensity and textural features, which are completely ignored in this study, in order to focus exclusively on the structural aspects of target object categories. The obtained classifiers, and the primitives selected by our method, aim to construct an intermediate step between the low-level image characteristics and high-level semantic concepts.

The exploration and use of more explicit spatial relations, as *in the center*, *around*, *between*, between object parts is an important perspective of our studies.

7 Acknowledgments

We would like to thank French Space Agency - CNES, Toulouse, and especially Gilbert Pauc and Jordi Inglada for their support.

References

Agarwal, S., Roth, D., 2002. Learning a Sparse Representation for Object Detection. ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV.

Bardinet, E., Vidal, S., Arroyo, S., Malandain, G., Blanca, N., 2000. Structural object matching, DECSAI-000303, Spain.

Biederman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychol Rev, American Psychological Association, Inc. 2, 115–147.

Biederman, I., 2007. Recent psychophysical and neural research in shape recognition. Object Recognition, Attention, and Action, Editor: N. Osaka, I. Rentschler, I. Biederman, Springer, pg. 71–88

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell., IEEE Computer Society, 679–698.

Cao, L., Kobayashi, Y., Kuno, Y., 2009. Spatial Relation Model for Object Recognition in Human-Robot Interaction., ICIC (1) 2009: 574-584.

Cheng, Y., 1995. Mean Shift, Mode Seeking, and Clustering. IEEE Trans. Pattern Anal. Mach. Intell., IEEE Computer Society. 17-8, 790–799.

Colliot, O., Camara, 0., Bloch, I., 2006. Integration of Fuzzy Spatial Relations in Deformable Models - Application to Brain MRI Segmentation. Pattern Recognition. 39, 1401–1414.

Comaniciu, D., Meer, P., 2002. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24-5, 603–619.

Cordella, L. P., Foggia, P., Sansone, C., Vento, M., 2002. Learning structural shape descriptions from examples. Pattern Recognition Letters, 23-12, 1427–1437.

Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. ECCV International Workshop on Statistical Learning in Computer Vision. 1–22.

Devernay, F., 1995. A Non-Maxima Suppression Method for Edge Detection with Sub-Pixel Accuracy, INRIA, France, RR-2724, 20 p.

Di Ruberto, C., 2004. Recognition of shapes by attributed skeletal graphs. Pattern Recognition. 37-1, 21–31.

Dorkó, G., Schmid, C., 2003. Selection of Scale-Invariant Parts for Object Class Recognition. ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision.

Douglas, D.H., Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Canadian Cartographer. 10, 112–122.

Erus, G., Lomenie, N., 2005. Automatic learning of structural models of cartographic objects. IAPR, Graph-Based Representations in Pattern Recognition, Springer, Lectures Notes in Computer Science. 3434, 273–280.

Erus, G., Lomenie, N., 2007. Classification of Structural Cartographic Objects Using Edge-Based Features. ISVC07, Lake Tahoe, USA. 385–392.

Felzenszwalb, P. F., Huttenlocher, D. P., 2005. Pictorial Structures for Object Recognition. Int. J. Comput. Vision. 61–1, 55–79.

Fergus, R., Perona, P., Zisserman, A., 2003. Object class recognition by unsupervised scale-invariant learning. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 264–271.

Ferrari, V., Fevrier L., Jurie, F., Schmid, C., 2008. Groups of Adjacent Contour Segments for Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 30–1, 36–51.

Fischler, M. A., Elschlager, R. A., 1973. The Representation and Matching of Pictorial Structures. IEEE Trans. Comput. 22–1, 67–92.

Heisele, B., Serre, T., Pontil, M., Vetter, T., Poggio, T., 2001. Categorization by Learning and Combining Object Parts. MIT Press. 1239–1245.

Hong, P., Huang, T. S., 2004. Spatial pattern discovery by learning a probabilistic parametric model from multiple attributed relational graphs. Discrete Appl. Math. 139 1-3, 113–135.

Hudelot, C., 2005. Towards a Cognitive Vision Platform for Semantic Image Interpretation, Application to the Recognition of Biological Organisms. Universit de Nice Sophia Antipolis - INRIA.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. PandRS. 62–3, 236-248.

Iqbal, Q., Aggarwal, J.K., 2002. Retrieval by Classification of Images Containing Large Manmade Objects Using Perceptual Grouping. Pattern Recognition Journal. 35–7, pg. 1463–1479.

Leibe, B., Leonardis, A., Schiele, B., 2004. Combined object categorization and segmentation with an implicit shape model. ECCV'04 Workshop on Statistical Learning in Computer Vision. May.

Li, F., Perona, P., 2005. A Bayesian Hierarchical Model for Learning Natural Scene Categories. CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Volume 2, 524–531.

Lorette, A., Descombes, X., Zerubia, J., 2000. Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area Extraction from Satellite Images. International Journal of Computer Vision. 36–3, 221-236.

Mayer, H., Hinz, S., Bacher, U., Baltsavias, E., 2006. A Test of Automatic Road Extraction Approaches, PCV06.

Opelt, A., Pinz, A., Zisserman, A., 2006. A Boundary-Fragment-Model for Object Detection. Proceedings of the European Conference on Computer Vision.

Petrakis, E. G. M., Faloutsos, C., 1995. Similarity Searching in Large Image Databases, Department of Computer Science, University of Maryland.

Sangineto, E., 2003. An abstract representation of geometric knowledge for object classification. Pattern Recogn. Letters. 24, 9-10, 1241–1250.

Shao, Z., Kittler, J., 1999. Shape representation and recognition based on invariant unary and binary relations. Image Vision Comput. 17, 5-6, 429-444.

Shotton, J., Blake, A., Cipolia, R., 2007. Multi-Scale Categorical Object Recognition Using Contour Fragments. IEEE Transactions on Pattern Analysis and Machine Intelligence.

D. M. J. Tax, 2001. One-class classification; Concept-learning in the absence of counter-examples. PhD Thesis. Delft University of Technology, Delft, The Netherlands.

Trias-Sanz, R., Loménie, N., 2003. Automatic bridge detection in high-resolution satellite images, 3rd International Conference on Computer Vision Systems (ICVS'03), LNCS - Springer, 172-181.

Tupin, F., Maitre, F., Mangin, J.-F., Nicolas, J. M., Pechersky, E., 1998. Detection of Linear Features in SAR Images: Application to the Road Network Extraction. IEEE Trans. Geosci. Remote Sensing. 36–2, 434–453.

Viola, P., Jones, M. J., 2004. Robust Real-Time Face Detection. Int. J. Comput. Vision, Kluwer Academic Publishers. 57–2, 137–154.

Wang, Q. and Seabra-Lopes, L. and Tax, D.M.J., 2004. Visual object recognition through one-class learning. Proceedings of ICIAR'04, vol. I, pg. 463–470.

Weber, M., Welling, M., Perona, P., 2000. Unsupervised Learning of Models for Recognition. ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part I, Springer-Verlag. 18–32.

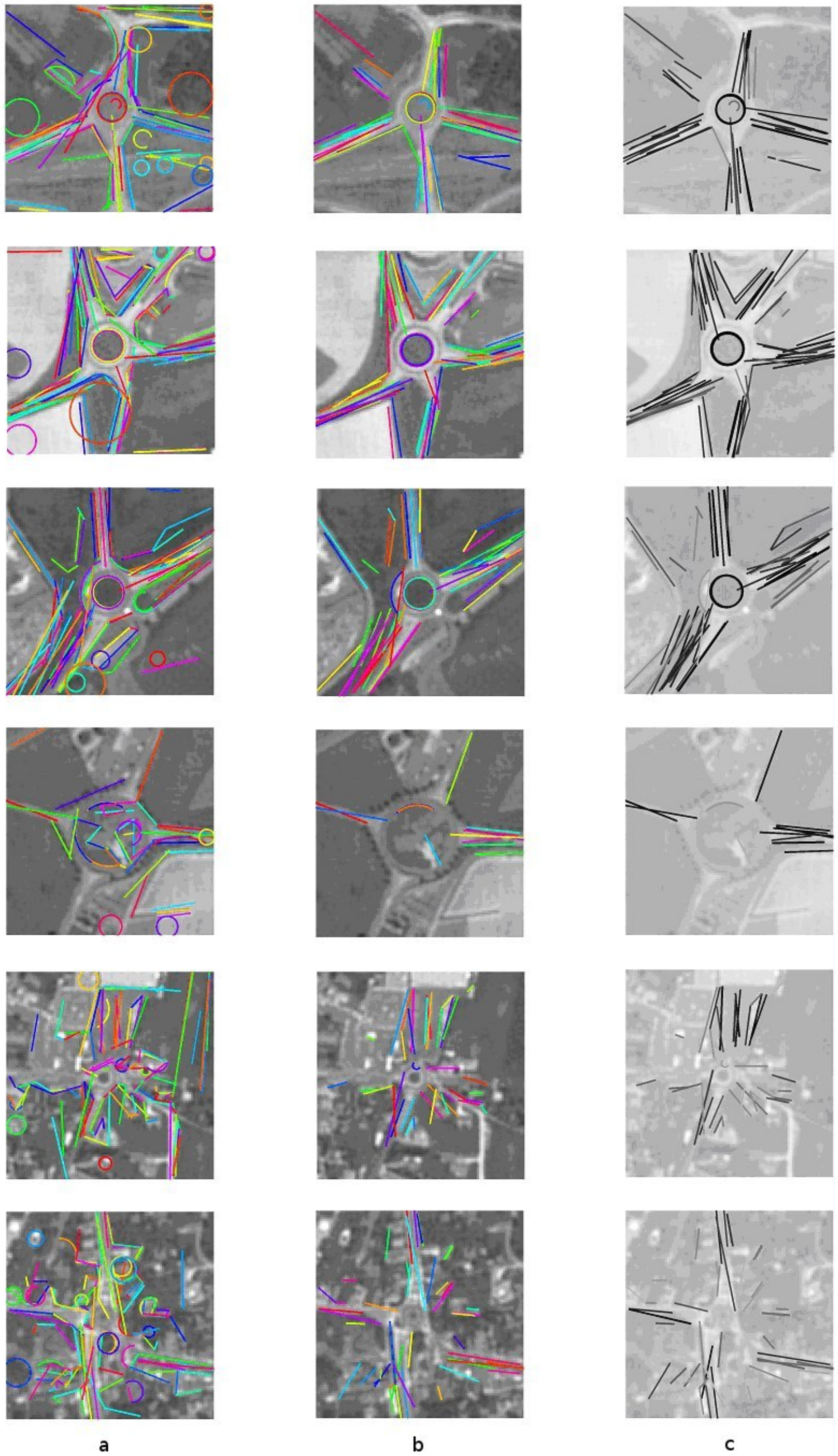


Fig. 9. The 3 best ranked, and the 3 worst ranked roundabout images in $\mathcal{SMC}_{\mathcal{RA}}$, (a). All detected primitives, (b). Selected primitives, (c). represented by a gray level corresponding to their membership score.

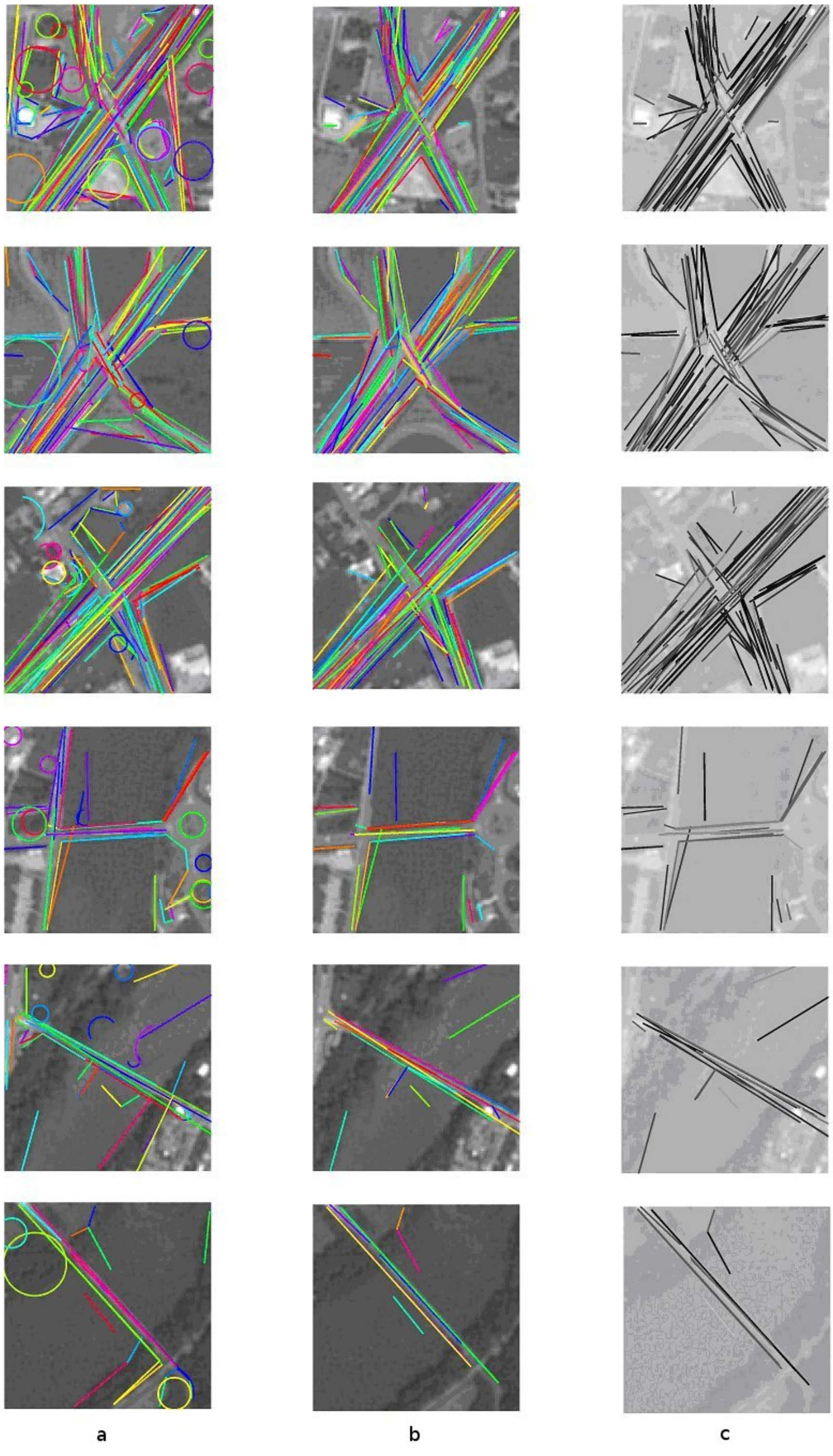


Fig. 10. The 3 best ranked, and the 3 worst ranked bridge images in $\mathcal{SMC}_{\mathcal{BR}}$, (a). All detected primitives, (b). Selected primitives, (c). represented by a gray level corresponding to their membership score.